VERFAHREN UND VOORICHTUNG ZUR BEARBEITUNG EINES SPRACHSIGNALS FÜR DIE ROBUSTE SPRACHERKENNUNG

Die Erfindung betrifft ein Verfahren und eine Vorrichtung zur Bearbeitung eines Sprachsignals, welches Rauschen aufweist, für eine anschließende Spracherkennung.

Spracherkennung wird in zunehmendem Maße eingesetzt, um die Bedienung von elektrischen Geräten zu erleichtern.
Um eine Spracherkennung zu ermöglichen, muss ein sogenanntes akustisches Modell erstellt werden. Dazu werden Sprachkommandos trainiert, was beispielsweise - für den Fall einer sprecherunabhängigen Spracherkennung -schon werkseitig erfolgen kann. Unter Training versteht man dabei, dass auf der Basis von mehrfachem Sprechen eines Sprachkommandos sogenannte, das Sprachkommando beschreibende, Merkmalsvektoren erstellt werden. Diese Merkmalsvektoren (die auch Prototypen genannt werden) werden dann in dem akustischen Modell, beispielsweise einem sogenannten HMM (Hidden Markov Modell) gesammelt.
Das akustische Modell dient dazu einer gegebenen Folge von aus dem Vokabular ausgewählten Sprachkommandos bzw. Wörtern die Wahrscheinlichkeit der beobachteten Merkmalsvektoren (während der Erkennung) zu ermitteln.

Zur Spracherkennung bzw. Erkennung der fließenden Sprache wird neben einem akustischen Modell auch ein sogenanntes Sprachmodell benutzt, das die Wahrscheinlichkeit des Aufeinanderfolgens einzelner Wörter in der zu erkennenden Sprache angibt.

Ziel von derzeitigen Verbesserungen bei der Spracherkennung ist es, nach und nach bessere Spracherkennungsraten zu erzielen, d.h. die Wahrscheinlichkeit zu erhöhen, dass ein von einem Benutzer des mobilen Kommunikationsgeräts gesprochenes Wort oder Sprachkommando auch als dieses erkannt

wird.
Da diese Spracherkennung vielseitig eingesetzt wird, erfolgt die Benutzung auch in Umgebungen, die durch Geräusch gestört sind. In diesem Fall sinken die Spracherkennungsraten drastisch, da die im akustischen Modell, beispielsweise dem HMM befindlichen Merkmalsvektoren auf Basis von reiner, d.h. nicht mit Rauschen behafteter Sprache erstellt wurden. Dies führt zu einer unbefriedigenden Spracherkennung in lauten Umgebungen, wie etwa auf der Straße, in viel besuchten Gebäuden oder auch im Auto.

Ausgehend von diesem Stand der Technik ist es Aufgabe der Erfindung, eine Möglichkeit zu schaffen, Spracherkennung auch in geräuschbehafteten Umgebungen mit einer hohen Spracherkennungsrate durchzuführen.

Diese Aufgabe wird durch die unabhängigen Ansprüche gelöst. Vorteilhafte Weiterbildungen sind Gegenstand der abhängigen Ansprüche.

Es ist Kern der Erfindung, dass eine Verarbeitung des Sprachsignals stattfindet, bevor dieses beispielsweise einer Spracherkennung zugeführt wird. Im Rahmen dieser Verarbeitung erfährt das Sprachsignal eine Geräuschunterdrückung. Anschließend wird das Sprachsignal hinsichtlich seine Signallevels bzw. Signalpegels normiert. Das Sprachsignal umfasst hierbei ein oder mehrere Sprachkommandos.

Dies hat den Vorteil, dass die Spracherkennungsraten für ein Sprachkommando bei einem derart vorverarbeiteten Sprachsignal mit geräuschbehafteter Sprache signifikant höher sind als bei einer herkömmlich Spracherkennung mit geräuschbehafteten Sprachsignalen.

Optional kann das Sprachsignal nach der Geräuschunterdrückung auch einer Einheit zur Bestimmung der Sprachaktivität zugeführt werden. Aufgrund dieses geräuschreduzierten

Sprachsignals wird dann festgestellt ob Sprache oder eine Sprachpause vorliegt. In Abhängigkeit davon wird der Normierungsfaktor für eine Signallevelnormierung festgelegt. Insbesondere kann der Normierungsfaktor so festgelegt werden, dass Sprachpausen stärker unterdrückt werden. Damit wird der Unterschied zwischen Sprachsignalabschnitten, in denen Sprache vorliegt und solchen, in denen keine vorliegt (Sprachpausen), noch deutlicher. Dies erleichtert eine Spracherkennung.

Ein Verfahren mit den oben beschriebenen Merkmalen kann auch bei sogenannten verteilten Spracherkennungssystemen angewendet werden. Ein verteiltes Spracherkennungssystem ist dadurch gekennzeichnet, dass nicht alle Schritte im Rahmen der Spracherkennung in derselben Komponente durchgeführt werden. Es ist also mehr als eine Komponenten erforderlich. Beispielsweise kann es sich bei einer Komponente um ein Kommunikationsgerät und bei einer weiteren Komponente um ein Element eines Kommunikationsnetzwerkes handeln. Hierbei findet beispielsweise die Sprachsignalerfassung bei einem als Mobilstation ausgestalteten Kommunikationsgerät statt, die eigentliche Spracherkennung dagegen in dem Kommunikationsnetzwerk-Element netze-seitig.

Dieses Verfahren lässt sich sowohl bei der Spracherkennung anwenden, als auch bereits bei der Erstellung des akustischen Modells, beispielsweise eines HMM's. Eine Anwendung bereits bei der Erstellung von akustischen Modellen zeigt in Zusammenhang mit einer Spracherkennung, die auf einem erfindungsgemäß vorverarbeiteten Signal basiert, eine weitere Erhöhung der Spracherkennungsrate.

Weitere Vorteile werden anhand ausgewählter Ausführungsbeispiele dargestellt, die auch in den Figuren abgebildet sind.

Es zeigen:

Fig.1: Ein Histogramm, in dem Sprachsignale, die ein oder mehrere Sprachkommandos enthalten, gegenüber ihrem Signallevel aufgetragen sind, für den Fall eines Trainings zur Erstellung eines akustischen Modells;

Fig.2: Ein Histogramm von Sprachsignalen gegenüber ihrem Signallevel für den Fall einer Spracherkennung;

Fig.3: Eine schematische Ausgestaltung einer erfindungsgemäßen Verarbeitung;

Fig.4: Ein Histogramm, in dem das geräuschreduzierte und sprachlevelnormierte Sprachsignal gegen den Sprachsignallevel aufgetragen ist;

Fig. 5 Ein Histogramm, in dem das geräuschreduzierte Sprachsignal gegenüber dem Signallevel aufgetragen ist;

Fig. 6 Ein Histogramm, in dem das Sprachsignal im Training erfindungsgemäß vorverarbeitet wird;

Fig. 7 Das Schema einer verteilten Sprachverarbeitung;

Fig. 8 Ein elektrisches Gerät, welches im Rahmen einer verteilten Sprachverarbeitung einsetzbar ist.

In Fig. 8 ist ein als Mobiltelefon bzw. Mobilstation MS ausgebildetes elektrisches Gerät dargestellt. Es verfügt über ein Mikrofon M zur Aufnahme von Sprachkommandos enthaltender Sprachsignale, eine Prozessoreinheit CPU zur Verarbeitung der Sprachsignale und eine Funkschnittstelle FS zum Übermitteln von Daten, beispielsweise verarbeiteten Sprachsignalen.

Das elektrische Gerät kann allein oder im Zusammenhang mit anderen Komponenten eine Spracherkennung bezüglich des aufgenommenen bzw. erfassten Sprachkommandos realisieren.

Es sollen nun zunächst eingehende Untersuchungen dargestellt werden, die zur Erfindung geführt haben:

In Fig. 1 ist ein Histogramm zu sehen, in dem Sprachsignale, welche eines oder mehrere Sprachkommandos enthalten, bezüglich ihres Signallevels L sortiert wurden und diese Häufigkeit H gegenüber dem Signallevel bzw. -pegel L aufgetragen wurde. Dabei enthält ein Sprachsignal S, wie es z.B. in den folgenden Figuren bezeichnet wird, ein oder mehrere Sprachkommandos. Zur Vereinfachung sei im Folgenden angenommen, dass das Sprachsignal ein Sprachkommando enthalte. Ein Sprachkommando kann beispielsweise bei einem als Mobiltelefon ausgestalteten elektrischen Gerät durch die Aufforderung "Anruf" sowie optional einem bestimmten Namen gebildet werden. Ein Sprachkommando muss bei einer Spracherkennung trainiert werden, d.h. auf Basis eines oftmaligen Sprechens des Sprachkommandos wird ein Merkmalsvektor oder werden mehrere, d.h. mehr als ein, Merkmalsvektoren erstellt. Dieses Training findet im Rahmen der Erstellung des akustischen Modells, beispielsweise des HMM's statt, welches bereits herstellerseitig erfolgt. Diese Merkmalsvektoren werden später zur Spracherkennung herangezogen.

Das Training von Sprachkommandos, welches zur Erstellung von Merkmalsvektoren dient, wird auf einem festgelegten Signallevel bzw. Lautstärkepegel durchgeführt ("Single Level Training"). Um den dynamischen Bereich des AD-Wandlers zum Umwandeln des Sprachsignals in ein digitales Signal optimal auszunutzen, wird vorzugsweise bei -26 dB gearbeitet. Die Festlegung auf Dezibel (dB) ergibt sich aus den für den Signallevel zur Verfügung stehenden Bits. So würde 0 dB einen Überlauf bedeuten (also ein Überschreiten der maximalen Lautstärke bzw. des maximalen Pegels). Alternativ kann anstelle eines "Single Level Trainings" auch ein Training auf mehreren Signallevels, beispielsweise bei -16, -26 und -36

dB durchgeführt werden.

In Fig. 1 ist hierbei die Häufigkeitsverteilung des Sprachlevels bei einem Sprachkommando für ein Training zu sehen.

Es ergeben sich für ein Sprachkommando ein mittlerer Signalwert X_{mean} sowie eine gewisse Verteilung der Levels des Sprachsignals. Dies kann als eine Gauss-Funktion mit dem mittleren Signallevel X_{mean} und einer Varianz σ dargestellt werden.

Nachdem in Fig. 1 die Verteilung der Sprachkommandos für eine Trainingssituation zu sehen ist, ist in Fig. 2, welche wiederum die Häufigkeit H gegenüber dem Signallevel L entsprechend Fig. 1 angibt die Situation bei einer Spracherkennung dargestellt: Es ist hier das Sprachsignal S' mit einem oder mehreren Sprachkommandos, wie es in den nachfolgenden Figuren bezeichnet wird, hinsichtlich seines Signallevels L sortiert und die Häufigkeit H aufgetragen. Aufgrund von Umgebungseinflüssen ergibt sich auch nach einer bereits angewendeten Geräuschunterdrückung NR (vgl. Fig. 3) eine gegenüber der Trainingssituation in Fig. 1 verschobene Verteilung mit einem neuen, gegenüber dem Mittelwert X_{mean} im Training verschobenen mittleren Signallevel x_{mean}.

Es hat sich in Untersuchungen erwiesen, dass die Spracherkennungsrate aufgrund dieses verschobenen mittleren Signallevels x_{mean} drastisch zurückgeht.

Dies ist aus der nachfolgenden Tabelle 1 zu ersehen:

Tabelle 1: Training mit reiner ("clean") Sprache verschiedener Lautstärkestufen bzw. Signallevel (Multi-Level).
Die Spracherkennungsraten beziehen sich auf Testsprache, die auf die Signallevel -16, -26, -36 dB normalisiert wurde.

Testsprache: Signallevel	Worterkennungsraten [%]							
	Subway		Babble		Car		Exhibition	
	Clean	5 dB	Clean	5 dB	Clean	5 dB	Clean	5 dB
-16 dB	98.83	80.14	98.79	66.99	98.72	88.01	99.11	79.76
-26 dB	99.14	85.66	99.15	76.66	99.19	91.35	99.35	85.00
-36 dB	99.39	85.05	99.21	82.41	99.28	89.41	99.57	85.47

In Tabelle 1 ist die Spracherkennungsrate bzw. Worterkennungsrate für verschiedene Geräuschumgebungen aufgeführt, wobei ein Training mit geräuschfreier Sprache ("Clean Speech") verschiedener Lautstärke stattgefunden hat. Die Testsprache, also das Sprachsignal aus Fig. 1 wurde auf drei unterschiedliche Levels bzw. Pegeln bei -16 dB, -26 dB und -36 dB normiert. Für diese unterschiedlichen Testsprachenergielevel sind die Spracherkennungsraten für unterschiedliche Arten von Geräuschen mit einem Geräuschpegel von 5 dB aufgezeigt. Bei den unterschiedlichen Geräuschen handelt es sich um typische Umgebungsgeräusche wie etwa U-Bahn bzw. "subway", sogenanntes Babble Noise, d.h. z.B. eine Cafeteria-Umgebung mit Sprache und anderen Geräuschen, das Hintergrundgeräusch in einem Auto bzw. "car", sowie eine Ausstellungsumgebung bzw. "exhibition", (d.h. ähnlich wie Babble Noise nur schlimmer evtl. mit Durchsagen, Musik usw.). Aus der Tabelle 1 ist ersichtlich, dass die Spracherkennung bei geräuschfreier Sprache weitgehend unbeeinflusst ist von Variationen im Testspracheenergielevel. Allerdings ist für geräuschbehaftete Sprache signifikanter Abfall der Spracherkennung zu erkennen. Zur Spracherkennung wurde hierbei die weiter unten beschriebene terminalbasierte Vorverarbeitung AFE, die zur Erstellung der Merkmalsvektoren dient, herangezogen.

Bei den in Tabelle 1 untersuchten Spracherkennungsraten - die gleichwohl nicht befriedigend sind- ist die Situation dennoch gegenüber einer Spracherkennung basierend auf einem Training mit nur einer Lautstärkenstufe wesentlich verbessert.

In anderen Worten, der Effekt, den ein Umgebungsgeräusch auf ein akustisches Modell hat, das auf Basis nur einer Lautstärke der Trainingssprache erstellt wurde, ist noch deutlicher verschlechternd.

Dies hat zu den im folgenden dargestellten erfindungsgemäßen Verbesserungen geführt:

In Fig. 3 ist nun der Ablauf gemäß einer Ausführungsform der Erfindung dargestellt. Das Sprachkommando bzw. Sprachsignal S, z.B. ein von einem Menschen gesprochenes Wort erfährt eine Geräuschunterdrückung NR. Nach dieser Geräuschunterdrückung NR liegt ein geräuschunterdrücktes Sprachsignal S' vor.

Das geräuschreduzierte Sprachsignal S' wird anschließend einer Signallevelnormierung bzw. Normierung des Signalwertes SLN unterzogen. Diese Normierung dient zur Herstellung eines Signalwertes, der mit dem mittleren Signalwert, der in Fig. 1 mit X_{mean} gekennzeichnet ist, vergleichbar ist. Es hat sich herausgestellt, dass bei vergleichbaren Signalmittelwerten höhere Spracherkennungsraten erzielt werden. Das heißt, dass durch diese Verschiebung des Signalwertes die Spracherkennungsrate bereits erhöht wird.

Im Anschluss an die Signalwertnormierung SLN liegt ein normiertes und geräuschreduziertes Sprachsignal S'' vor. Dies kann im Folgenden z.B. bei einer Spracherkennung SR mit einer höheren Spracherkennungsrate auch bei einer ursprünglich mit Rauschen behafteten Testsprache, verwendet werden.

Optional wird das geräuschreduzierte Signal S' aufgespalten und fließt neben der Signalwertnormierung SLN auch einer Sprachaktivitätsbestimmungseinheit bzw. "Voice Activity Detection" VAD zu. In Abhängigkeit davon, ob Sprache oder eine Sprachpause vorliegt, der Normierungswert, mit dem das geräuschreduzierte Sprachsignal S' normiert wird, eingestellt

werden. Beispielsweise kann in Sprachpausen ein kleinerer multiplikativer Normierungsfaktor verwendet werden, wodurch der Signallevel des geräuschreduzierten Sprachsignals S' in Sprachpausen stärker reduziert wird, als während des Vorliegens von Sprache. Damit ist eine stärkere Unterscheidung zwischen Sprache, also z.B. einzelnen Sprachkommandos, und Sprachpausen möglich, was eine nachgeschaltete Spracherkennung hinsichtlich der Spracherkennungsrate weiter deutlich verbessert.

Weiterhin ist es vorgesehen, den Normierungsfaktor nicht nur zwischen Sprachpausen und Sprachabschnitten zu verändern, sondern auch innerhalb eines Wortes für unterschiedliche Sprachabschnitte zu variieren. Auch dadurch kann die Spracherkennung verbessert werden, da einige Sprachabschnitte aufgrund der in ihnen enthaltenen Phoneme einen sehr hohen Signallevel, beispielsweise bei Plosivlauten (z.B. p), aufweisen, während andere eher inhärent leise sind.

Für die Signallevelnormierung werden unterschiedliche Methoden herangezogen, beispielsweise eine Echt-Zeit-Energie-Normalisierung, wie sie im Artikel "Robust Endpoint Detection and Energy Normalisation for Real-Time Speech and Speaker recognition" von Qi Li et al. in IEEE Transactions on Speech and Audio Processing Vol. 10, No. 3, März 2002 im Abschnitt C (S. 149-150) beschrieben wird. Im Rahmen der ITU wurde weiterhin eine Signallevelnormierungsmethode beschrieben, die unter ITU-T, ``SVP56: The Speech Voltmeter'', in Software Tool Library 2000 User's Manual, Seiten 151-161, Genf, Schweiz, Dezember 2000 zu finden ist. Die dort beschriebene Normierung arbeitet "off-line" bzw. in einem sogenannten "Batch-Modus", d.h. nicht zeitgleich bzw. zeitnahe mit der Spracherfassung.

Für die Geräuschreduktion bzw. Geräuschunterdrückung NR (vgl. Fig.3) sind ebenfalls verschiedene bekannte Methoden vorgesehen, beispielsweise im Frequenzraum operierende

Methoden. Eine solche Methode ist in "Computationally efficient speech enhancement using RLS and psycho-acoustic motivated algorithm" von Ch. Beaugeant et al. in Proceedings of 6th World Multi-conference on Systemics, Cybernetics and Informatics, Orlando 2002 beschrieben. Das dort beschrieben System basiert auf einem Analyse-durch-Synthese System, bei dem rahmenweise rekursiv die das (reine) Sprachsignal und das Rauschsignal beschreibende Parameter extrahiert werden (vgl. dort Abschnitt 2 "Noise Reduction in the Frequency Domain", Abschnitt 3 "Recursive implementation of the least square algorithm"). Das so erhaltene reine Sprachsignal wird weiterhin gewichtet (Vgl. Abschnitt 4 "Practical RLS Weighting Rule") und eine Schätzung der Leistung des Rauschsignals erfolgt (Vgl. Abschnitt 5 "Noise Power Estimation"). Optional kann eine Verfeinerung des erhaltenen Resultats mittels psychoakustisch motivierter Methoden erfolgen (Abschnitt 6: "Psychoacoustic motivated method"). Weitere Geräuschreduktionsmethoden, die gemäß einer Ausführungsform nach Fig. 3 herangezogen werden können sind beispielsweise in ETSI ES 202 0505 V1.1.1 vom Oktober 2002 in Abschnitt 5.1 ("Noise Reduction") beschrieben.

Ein in Bezug auf Geräuschunterdrückung NR und Signallevelnormierung SN unbearbeitetes Sprachsignal S liegt den Häufigkeitsverteilungen in den Fig. 1 (Trainingssituation) und 2 (Testsituation, d.h. für eine Spracherkennung) zugrunde. Das geräuschreduziert Sprachsignal S' liegt der Häufigkeitsverteilung in der Figur 5 zugrunde. Das geräuschreduzierte und signallevelnormierte Signal liegt den Verteilungen in den Figuren 4 (Testsituation) und 5 (Trainingssituation) zugrunde.

Die zugrundeliegende Idee des in Fig. 3 gezeigten, schematischen Ablaufes einer Sprachsignalverarbeitung zu einer nachgeordneten Spracherkennung ist in den Figuren 4 bis 6 dargestellt.

In Fig. 5 ist eine Häufigkeitsverteilung für ein geräuschreduziertes Sprachsignal S' dargestellt, wie es z.B. in Fig. 3 nach der Geräuschunterdrückung NR auftritt. Gegenüber Fig. 2, die sich z.B. auf die Häufigkeitsverteilung für ein in Fig. 3 dargestelltes Sprachsignal S bezieht, wurde also nach eine Geräuschunterdrückung NR durchgeführt.

Das Zentrum der Häufigkeitsverteilung dieses geräuschreduzierten Sprachsignals S' gegenüber dem Sprachlevel L befindet sich bei einem Mittelwert x_{mean} '. Die Verteilung hat eine breite σ'. Im Übergang zu Fig. 4 wird auf das in Fig. 5 dargestellte geräuschreduzierte Sprachsignal S' eine Signallevelnormierung SLN durchgeführt. Damit würde das der Verteilung in Fig. 4 zugrundeliegende Sprachsignal beispielsweise dem geräuschreduzierten und signallevelnormierten Sprachsignal S'' entsprechen. Eine Signallevelnormierung bringt den tatsächlichen Signallevel in Fig. 5, auf einen gewünschten Signallevel, beispielsweise den in Fig. 1 mit X_{mean} gekennzeichneten, im Training erzielten Signallevel. Weiterhin führt die Signallevelnormierung SLN dazu, dass die Verteilung schmaler wird, d.h. also dass σ'' kleiner ist als σ'. Dadurch kann der mittlere Signallevel x_{mean}'' in Fig. 4 leichter mit dem mittleren Signallevel X_{mean} in Fig. 1, welcher im Training erzielt wurde, zur Deckung gebracht werden. Dies führt zu höheren Spracherkennungsraten.

Im Zusammenhang mit Fig. 7 wird nun auf eine Anwendung des oben erläuterten für eine Spracherkennung eingegangen. Wie bereits eingangs dargelegt, kann die Spracherkennung in einer Komponente oder auf mehrere Komponenten verteilt stattfinden.

Beispielsweise können sich in einem elektrischen Gerät MS, welches als Mobilstation ausgebildet ist, Mittel zum Erfassen des Sprachsignal, z.B. das in Fig. 8 gezeigt Mikrofon M,

Mittel zur Geräuschunterdrückung NR und Mittel zur Signallevelnormierung SN befinden. Letztere können im Rahmen der Prozessoreinheit CPU realisiert werden. Damit kann die in Fig. 3 dargestellte Idee einer Sprachsignalverarbeitung gemäß einer Ausführungsform der Erfindung sowie die sich anschließende Spracherkennung in einem Mobilfunkgerät bzw. Mobilstation allein oder im Zusammenhang mit einem Element eines Kommunikationsnetzes implementiert werden.

Gemäß einer der Alternativen erfolgt die Spracherkennung SR (siehe Fig. 3) selbst netz-seitig. Dazu werden die aus einem Sprachsignal S'' erstellten Merkmalsvektoren über einen Kanal, insbesondere einen Funkkanal zu einer zentralen Einheit im Netz übertragen. Dort findet auf Basis der übertragenen Merkmalsvektoren dann die Spracherkennung auf Basis des insbesondere bereits werkseitig erstellten Modells statt. Werkseitig kann insbesondere bedeuten, dass das akustische Modell vom Netzbetreiber erstellt wird.

Insbesondere kann die vorgeschlagene Spracherkennung auf sprecherunabhängige Spracherkennung, wie sie im Rahmen des sogenannten Aurora Szenarios vorgenommen wird, angewendet werden.
Eine weitere Verbesserung ergibt sich, wenn Sprachkommandos bereits bei der werkseitigen Herstellung des akustischen Modells bzw. dem Training hinsichtlich ihres Signallevels normiert werden. Dadurch wird nämlich die Verteilung der Signallevel schmaler, wodurch eine noch bessere Übereinstimmung zwischen der in Fig. 4 gezeigten Verteilung und der im Training erzielten Verteilung erreicht wird. Eine solche Verteilung der Häufigkeit H gegenüber dem Signalpegel L bei einem Sprachkommando im Training, bei dem bereits eine Signallevelnormierung durchgeführt wurde, ist in Fig. 6 dargestellt. Der sich ergebende Trainings-Mittelwert x_{mean_neu} stimmt mit dem dem Mittelwert x_{mean}'' (Fig.4) der geräuschreduzierten und signallevelnormierten Sprachsignals S'' (Fig.3) überein. Wie bereits dargelegt ist eine

Übereinstimmung der Mittelwerte eines der Kriterien für eine hohe Spracherkennungsrate.Weiterhin ist die Breite der Verteilung in Fig. 6 sehr schmal, was es erleichtert, diese Verteilung mit der Verteilung in Fig. 4 zur Deckung zu bringen, d.h. auf den gleichen Signallevel zu bringen.

In Fig.7 ist eine verteilte Spracherkennung bzw. "Distributed Speech Recognition" (DSR)dargestellt. Eine verteilte Spracherkennung kann beispielsweise im Rahmen bereits erwähnten AURORA-Projekts der ETSI STQ (Speech Transmission Quality) Anwendung finden.

Bei einer verteilten Spracherkennung wird bei einer Einheit ein Sprachsignal, beispielsweise ein Sprachkommando erfasst und dieses Sprachsignal beschreibende Merkmalsvektoren erstellt. Diese Merkmalsvektoren werden zu einer anderen Einheit, beispielsweise einem Netzwerkserver übertragen. Dort werden die Merkmalsvektoren verarbeitet und auf Basis dieser Merkmalsvektoren eine Spracherkennung durchgeführt.

In Fig. 7 ist eine Mobilstation MS als erste Einheit bzw. Komponente und eine Netzwerkelement NE dargestellt.

Die Mobilstation MS, welche auch als Terminal bezeichnet wird, weist Mittel AFE zurterminalbasierte Vorverarbeitung, die zur Erstellung der Merkmalsvektoren dient, . Beispielsweise handelt es sich bei der Mobilstation MS um ein Mobilfunk-Endgerät, portablen Computern, oder ein beliebiges anderes mobiles Kommunikationsgerät. Bei dem Mittel AFE zur terminalbasierten Vorverarbeitung handelt es sich beispielsweise um das im Rahmen des AURORA-Projekts diskutierte "Advanced Front End".

Das Mittel AFE zur terminalbasierten Vorverarbeitung umfasst Mittel zur Standardbearbeitung von Sprachsignalen. Diese Standard-Sprachverarbeitung ist beispielsweise in der Spezifikation ETSI ES 202050 V1.1.1 vom Oktober 2002 in Bild

4.1 beschrieben. Auf Seiten der Mobilstation beinhaltet die Standard-Sprachverarbeitung eine Merkmalsextraktion mit den Schritten Geräuschreduktion, Signalform bzw. "Waveform-Processing", Cepstrum-Berechnung sowie einen verdeckten Ausgleich bzw. "Blind Equalization". Anschließend erfolgt einer Merkmalskompression und eine Vorbereitung der Übertragung. Diese Verarbeitung ist dem Fachmann bekannt, weshalb hier nicht näher darauf eingegangen wird.
Gemäß einer Ausgestaltung der Erfindung umfassen die Mittel AFE zur terminalbasierten Vorverarbeitung auch Mittel zur Signallevenormierung und Sprachaktivitätsdetektion, damit eine Vorverabeitung gemäß Fig. 3 realisiert wird.

Diese Mittel können in die Mittel AFE integriert oder alternativ als getrennte Komponente realisiert sein.

Über sich anschließende Mittel FC zur Merkmalsvektorkomprimierung terminalbasierte Vorverarbeitung AFE werden der eine oder die mehreren Merkmalsvektoren, welche aus dem Sprachkommando erstellt werden, zum Zwecke der Übertragung über einen Kanal CH komprimiert.

Die andere Einheit wird beispielsweise durch einen Netzwerkserver als Netzwerkelement NE gebildet. In diesem Netzwerkelement NS werden die Merkmalsvektoren über Mittel FDC zur Merkmalsvektordekompression wieder dekomprimiert.Weiterhin erfolgt über Mittel SSP erfolgt eine serverseitige Vorverarbeitung , um dann mit Mitteln SR zur Spracherkennung eine Spracherkennung auf Basis eines Hidden Markov Modells HMM durchzuführen.

Die Ergebnisse von erfindungsgemäßen Verbesserungen werden nun erläutert: Spracherkennungsraten für verschiedene Trainings der Sprachkommandos sowie verschiedene Sprachlevel bzw. Lautstärken, die zur Spracherkennung herangezogen werden (Testsprache) sind in den Tabellen 1 bis 2 dargestellt.

In Tabelle 2 sind nun die Spracherkennungsraten für unterschiedliche Energielevel der Testsprache gezeigt. Das Training fand auf einem Sprachenergielevel von -26 dB statt. Die Testsprache wurde einer Geräuschunterdrückung und Sprachlevelnormalisierung gemäß Fig. 3 unterzogen. Aus Tabelle 2 ist zu sehen, dass die Spracherkennungsraten für reine Sprache wiederum gleichbleibend hoch sind. Die wesentliche Verbesserung gegenüber dem bisherigen Spracherkennungsverfahren liegt darin, dass der in Tabelle 1 ersichtliche Unterschied in den Spracherkennungsraten für geräuschbehaftete Sprache (bei einem Signal zu Rauschen Verhältnis bzw. "Signal-to-Noise Ratio" von 5 dB) in Abhängigkeit vom Energielevel der Testsprache aufgehoben ist. Für die Spracherkennung wurde das weiter oben beschriebene "Advanced Front End" herangezogen.

Tabelle 2:

Test Speech Energy Levels	Word Recognition Rates [%]							
	Subway		Babble		Car		Exhibition	
	Clean	5 dB	Clean	5 dB	Clean	5 dB	Clean	5 dB
-16 dB	99.45	83.79	98.85	75.63	99.02	86.34	99.35	79.67
-26 dB	99.20	84.71	98.88	74.37	99.05	87.89	99.32	80.56
-36 dB	98.86	84.71	98.70	75.00	98.78	87.77	99.01	80.47

Patentansprüche

1. Verfahren zur Bearbeitung eines geräuschbehafteten Sprachsignals (S) für eine nachfolgende Spracherkennung (SR), wobei das Sprachsignal (S) zumindest ein Sprachkommando repräsentiert, mit folgenden Schritten:
a) Erfassen des geräuschbehafteten Sprachsignals (S);
b) Anwendung einer Geräuschunterdrückung (NR) auf das Sprachsignal (S) zur Generierung eines geräuschunterdrückten Sprachsignals (S');
c) Normieren des geräuschunterdrückten Sprachsignals (S') mittels eines Normierungsfaktors auf einen Soll-Signalwert zur Generierung eines geräuschunterdrückten, normierten Sprachsignals (S'').

2. Verfahren nach Anspruch 1, bei dem der Wert des Normierungsfaktors in Abhängigkeit von einer Sprachaktivität festgelegt wird.

3. Verfahren nach Anspruch 1 oder 2, bei dem die Sprachaktivität auf Basis des geräuschunterdrückten Sprachsignals ermittelt wird.

4. Verfahren nach einem der vorhergehenden Ansprüche mit folgendem weiteren Schritt:
d) Beschreiben des geräuschunterdrückten, normierten Sprachkommandos durch einen oder mehrere Merkmalsvektoren.

5. Verfahren nach Anspruch 4, bei dem der eine oder die mehreren Merkmalsvektoren zum Beschreiben des geräuschunterdrückten, normierten Sprachkommandos erstellt werden.

6. Verfahren nach einem der vorhergehenden Ansprüche mit folgendem weiteren Schritt:
e) Übermitteln eines den Merkmalsvektor oder die Merkmalsvektoren beschreibenden Signals.

7. Verfahren nach einem der vorhergehenden Ansprüche mit folgendem weiteren Schritt:
f) Durchführen einer Spracherkennung auf Basis des geräuschunterdrückten, normierten Sprachkommandos.

8. Verfahren nach Anspruch 6 oder 7, bei dem das Erfassen der Sprachsignals in Schritt a) und das Durchführen der Spracherkennung in Schritt f) örtlich getrennt durchgeführt werden.

9. Verfahren nach einem der vorhergehenden Ansprüche, bei dem eine Vorverabeitung (AFE) und eine Merkmalsvektorkomprimierung (FC) von Merkmalsvektoren, welche ein Sprachsignal beschreiben räumlich getrennt oder ortsgleich durchgeführt wird.

10. Verfahren zum Training eines Sprachkommandos in einem geräuschbehafteten Sprachsignal mit folgenden Schritten:
a') Erfassen des geräuschbehafteten Sprachsignals;
b') Anwendung einer Geräuschunterdrückung auf das Sprachsignal zur Generierung eines geräuschunterdrückten Sprachsignals;
c') Normieren des geräuschunterdrückten Sprachsignals mittels eines Normierungsfaktors auf einen Soll-Signalwert zur Generierung eines geräuschunterdrückten, normierten Sprachsignals.

11. Verfahren nach Anspruch 10, bei dem das Training zur Erstellung eines akustischen Modells, insbesondere eines HMM's dient.

12. Elektrisches Gerät (MS) mit einem Mikrofon (M) und einer Prozessoreinheit (CPU), welches zur Durchführung eines

Verfahrens nach Anspruch 1 bis 11 eingerichtet ist, insbesondere zur Durchführung der Schritte a, b und c).

13.Vorrichtung nach Anspruch 12 mit einer Einrichtung zur Erstellung von Merkmalsvektoren zur Beschreibung eines Sprachsignals.

14. Elektrisches Gerät nach Anspruch 12 oder 13, welches als Kommunikationsgerät, insbesondere Mobilstation, ausgestaltet ist, mit einer Sende/Empfangseinrichtung (FS) und einer Vorrichtung nach Anspruch 12 oder 13.

15. Kommunikationssystem mit einer Mobilstation nach Anspruch 14 und einem Kommunikationsnetz, in dem eine Spracherkennung durchgeführt wird.

FIG 1

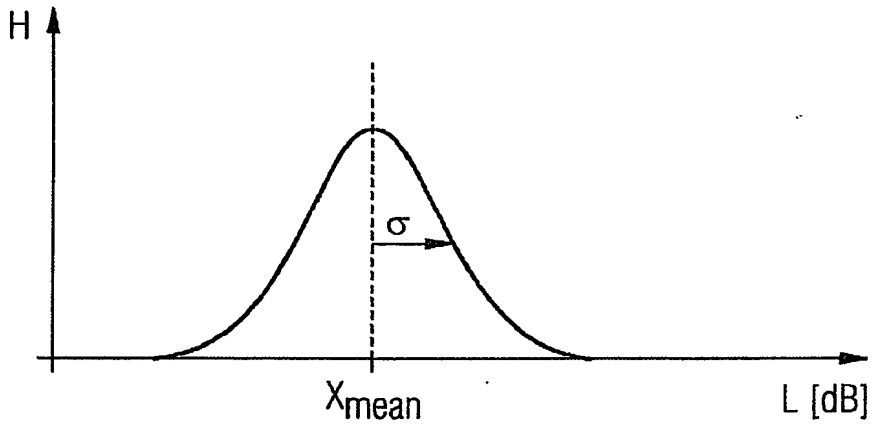


FIG 2

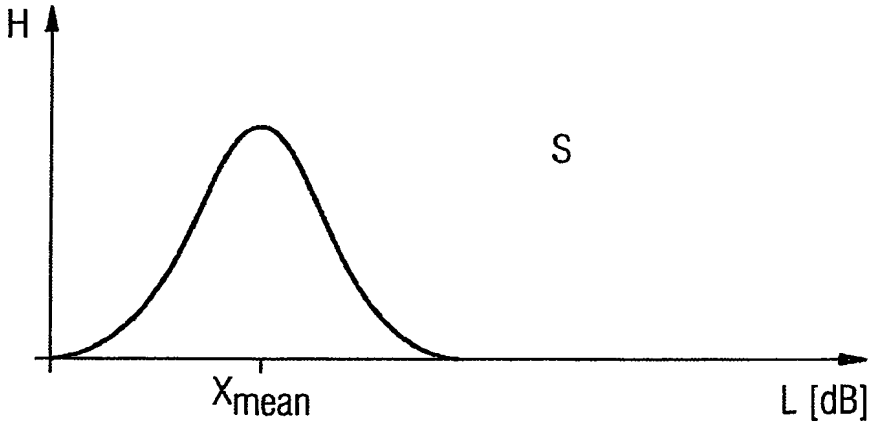


FIG 3

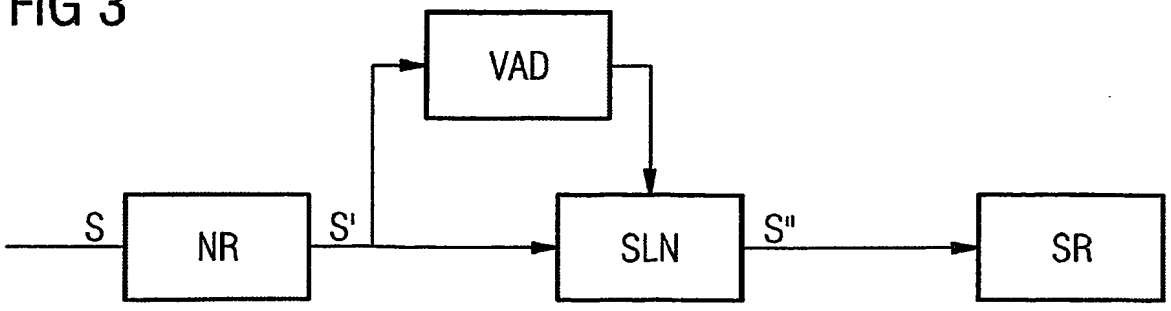


FIG 4

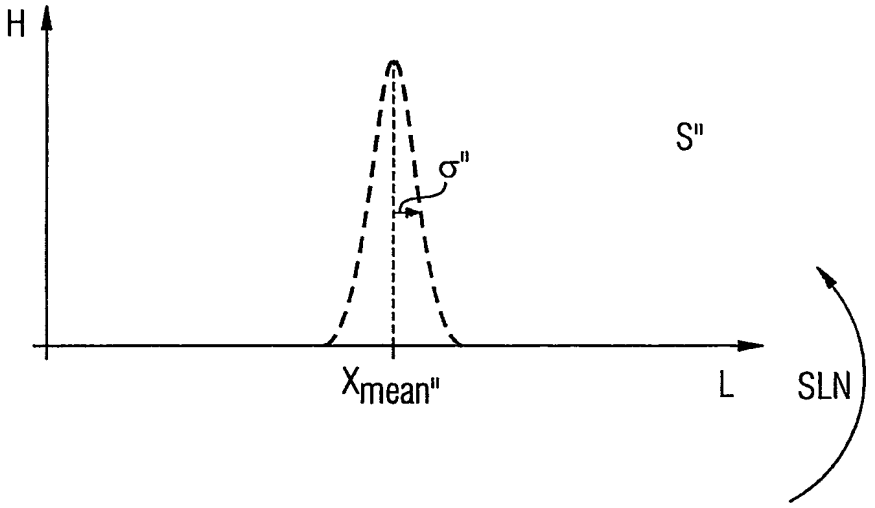


FIG 5

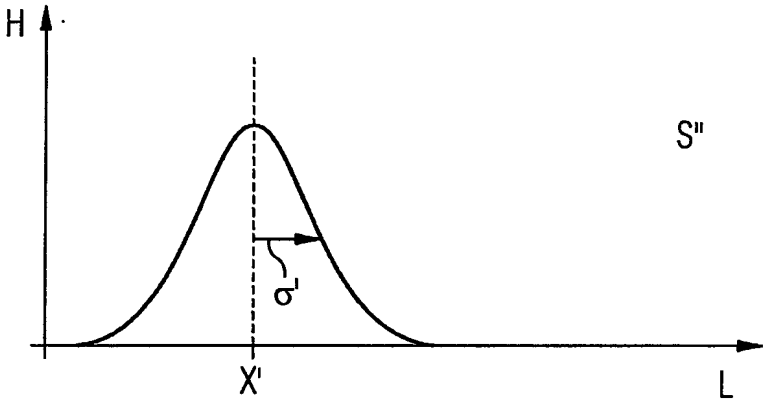


FIG 6

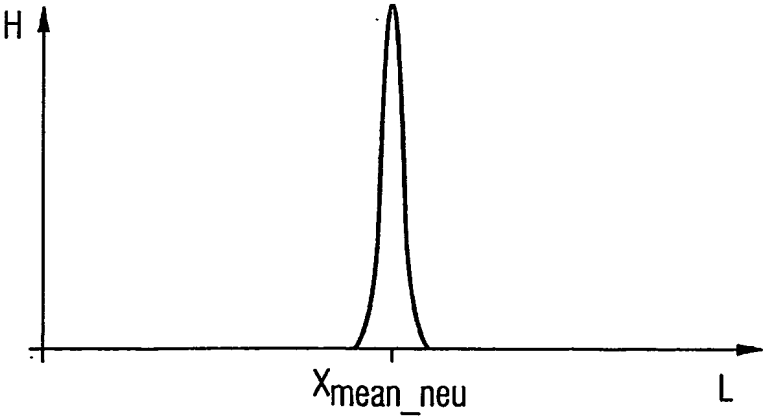


FIG 7

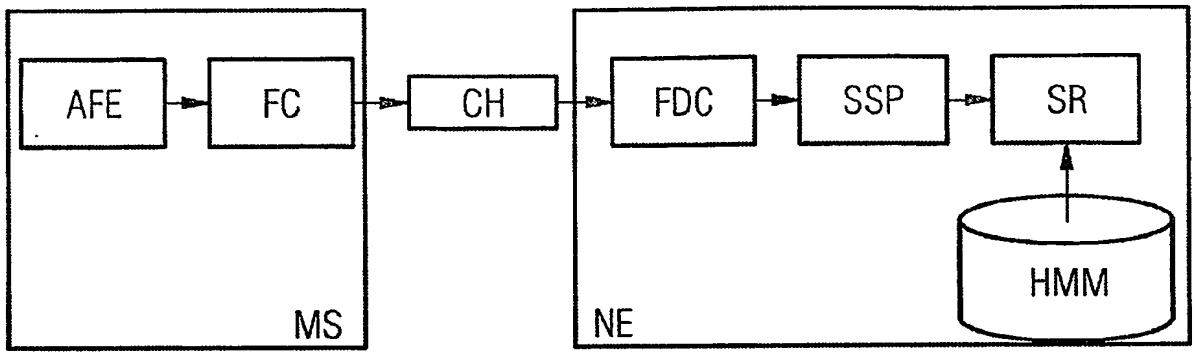
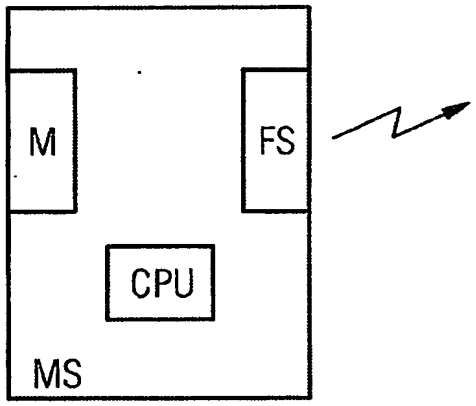


FIG 8



INTERNATIONAL SEARCH REPORT

International Application No
/EP2004/052427

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L21/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 6 098 040 A (PETERS STEVEN DOUGLAS ET AL) 1 August 2000 (2000-08-01) column 10, line 16 - line 31	1-15
X	DATABASE INSPEC 'Online! THE INSTITUTION OF ELECTRICAL ENGINEERS, STEVENAGE, GB; December 1998 (1998-12), SANG-MUN CHI ET AL: "The suppression of noise-induced speech distortions for speech recognition" XP008040656 Database accession no. 6261641 abstract -/—	1-15

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 December 2004	03/01/2005

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Burchett, S

1

Form PCT/ISA/210 (second sheet) (January 2004)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	& JOURNAL OF THE INSTITUTE OF ELECTRONICS ENGINEERS OF KOREA S INST. ELECTRON. ENG., vol. 35-S, no. 12, December 1998 (1998-12), pages 93-102, KOREA SOUTH KOREA ISSN: 1226-5837	
X	US 6 266 633 B1 (HIGGINS ALAN LAWRENCE ET AL) 24 July 2001 (2001-07-24) column 5, line 40 - line 66	1
A	FLORIAN HILGER AND HERMANN NEY: "NOISE LEVEL NORMALIZATION AND REFERENCE ADAPTATION FOR ROBUST SPEECH RECOGNITION" AUTOMATIC SPEECH RECOGNITION, CHALLENGES FOR THE NEW MILLENIUM, 18 September 2000 (2000-09-18), - 20 September 2000 (2000-09-20) pages 1-5, XP007005548 PARIS, FRANCE page 2, left-hand column, paragraph 2 - page 3, left-hand column, paragraph 4	1-15
A	RATHINAVELU CHENGALVARAYAN: "ROBUST ENERGY NORMALIZATION USING SPEECH/NONSPEECH DISCRIMINATOR FOR GERMAN CONNECTED DIGIT RECOGNITION" 6TH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. EUROSPEECH '99, vol. 1, 5 September 1999 (1999-09-05), - 9 September 1999 (1999-09-09) pages 61-64, XP007000915 BUDAPEST, HUNGARY page 62, right-hand column, paragraph 2.3	1-15

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/EP2004/052427

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
US 6098040	A	01-08-2000	NONE	
US 6266633	B1	24-07-2001	NONE	

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
[redacted]/EP2004/052427

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 G10L21/02

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G10L

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der Internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
EPO-Internal, INSPEC, WPI Data

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	US 6 098 040 A (PETERS STEVEN DOUGLAS ET AL) 1. August 2000 (2000-08-01) Spalte 10, Zeile 16 - Zeile 31	1-15
X	DATABASE INSPEC 'Online! THE INSTITUTION OF ELECTRICAL ENGINEERS, STEVENAGE, GB; Dezember 1998 (1998-12), SANG-MUN CHI ET AL: "The suppression of noise-induced speech distortions for speech recognition" XP008040656 Database accession no. 6261641 Zusammenfassung -/--	1-15

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[X] Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" älteres Dokument, das jedoch erst am oder nach dem Internationalen Anmeldedatum veröffentlicht worden ist
"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem Internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
"T" Spätere Veröffentlichung, die nach dem Internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der Internationalen Recherche	Absendedatum des Internationalen Recherchenberichts
21. Dezember 2004	03/01/2005

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	Burchett, S

1

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
	& JOURNAL OF THE INSTITUTE OF ELECTRONICS ENGINEERS OF KOREA S INST. ELECTRON. ENG., Bd. 35-S, Nr. 12, Dezember 1998 (1998-12), Seiten 93-102, KOREA SOUTH KOREA ISSN: 1226-5837	
X	US 6 266 633 B1 (HIGGINS ALAN LAWRENCE ET AL) 24. Juli 2001 (2001-07-24) Spalte 5, Zeile 40 - Zeile 66	1
A	FLORIAN HILGER AND HERMANN NEY: "NOISE LEVEL NORMALIZATION AND REFERENCE ADAPTATION FOR ROBUST SPEECH RECOGNITION" AUTOMATIC SPEECH RECOGNITION, CHALLENGES FOR THE NEW MILLENIUM, 18. September 2000 (2000-09-18), - 20. September 2000 (2000-09-20) Seiten 1-5, XP007005548 PARIS, FRANCE Seite 2, linke Spalte, Absatz 2 - Seite 3, linke Spalte, Absatz 4	1-15
A	RATHINAVELU CHENGALVARAYAN: "ROBUST ENERGY NORMALIZATION USING SPEECH/NONSPEECH DISCRIMINATOR FOR GERMAN CONNECTED DIGIT RECOGNITION" 6TH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. EUROSPEECH '99, Bd. 1, 5. September 1999 (1999-09-05), - 9. September 1999 (1999-09-09) Seiten 61-64, XP007000915 BUDAPEST, HUNGARY Seite 62, rechte Spalte, Absatz 2.3	1-15

1

INTERNATIONALER RECHERCHENBERICHT

Angaben zu Veröffentlichungen, die zur selben Patentfamilie gehören

Internationales Aktenzeichen
PCT/EP2004/052427

Im Recherchenbericht angeführtes Patentdokument		Datum der Veröffentlichung	Mitglied(er) der Patentfamilie	Datum der Veröffentlichung
US 6098040	A	01-08-2000	KEINE	
US 6266633	B1	24-07-2001	KEINE	